range, language preferences, marital status, geographic location (e.g., the city, state and country in which the user resides, and possibly also including additional information such as street address, zip code, and telephone area code), cultural background or preferences, or any subset of these. Alternatively, the geographic information can be inferred, for example, from the user's IP address, without having the user provide the geographic information explicitly. In particular, generally, one can map an IP address to an organization. If the organization is in one place (i.e. Stanford), then it is possible to infer the graphical location of the user searching from that IP address. The personal information **215** may also indicate whether the user is a member of in one or more defined groups (e.g., organizations, companies, associations, clubs, committees, and the like). The personal information **215** may also include psychographic information (e.g., personality trait information, or other personality descriptive information) either derived from other aspects of the user profile, or expressly provided by the user.

[**0040**] Compared with other types of personal information such as a user's favorite sports or movies that are often time varying, this personal information is more static and more difficult to infer from the user's search queries and search results, but maybe crucial in correctly interpreting certain queries submitted by the user. For example, if a user submits a query containing "Japanese restaurant", it is very likely that he may be searching for a local Japanese restaurant for dinner. Without knowing the user's geographical location, it is hard to order the search results so as to bring to the top those items that are most relevant to the user's true intention. In certain cases, however, it is possible to infer this information. For example, users often select results associated with a specific region corresponding to where they live.

[**0041**] Another potential source of information are expressed topics or category preferences **217**. The user profile can include a list of terms or topics that the user expressly indicates as being among the user's interests. The terms can be selected by the user from a predefined list or hierarchy of topics and terms, or provided by the entirely by the user. Each term or topic can be associated with a weight indicating a degree of importance to the user.

[**0042**] Another potential source of information for the user profile is information **219** derived from web pages and web sites associated with the user. First, a given user often accesses the system **100** from a relatively limited number of IP addresses and domains. The system **100** can automatically identify and access one or more websites associated with these IP addresses and extract information from them, such as their type (commercial, educational, organization, government, etc.), their geographic location, their size, and so forth. The system can further perform analyses of one or more of the pages on these sites (such as the home page), to extract relevant topics, key words, or other descriptive information.

[**0043**] Creating a user profile **230** from the various sources of user information is a multi-step process, which be divided into sub-processes. Each sub-process produces one type of user profile characterizing a user's interests or preferences from a particular perspective. They are:

[**0044**] a term-based profile **231**—this profile represents a user's search preferences with a plurality of terms, where each term is given a weight indicating the importance of the term to the user;

[**0045**] a category-based profile **233**—this profile correlates a user's search preferences with a set of categories, which may be organized in a hierarchal fashion, with each category being given a weight indicating the extent of correlation between the user's search preferences and the category; and

[**0046**] a link-based profile **235**—this profile identifies a plurality of links that are directly or indirectly related to the user's search preferences, with each link being given a weight indicating the relevance between the user's search preferences and the link.

[**0047**] In some embodiments, the user profile **230** includes only a subset of these profiles **231**, **233**, **235**, for example just one or two of these profiles. In one embodiment, the user profile **230** includes a term-based profile **231** and a category-based profile **233**, but not a link-based profile **235**.

[**0048**] In one embodiment, a user profile is created and stored on a server (e.g., user profile server **108**) associated with a search engine. The advantage of such deployment is that the user profile can be easily accessed by multiple computers, and that since the profile is stored on a server associated with (or part of) the search engine **104**, it can be easily used by the search engine **104** to personalize the search results. In another embodiment, the user profile can be created and stored on the user's client **118**. Creating and storing a user profile on the client not only reduces the computational and storage cost for the search engine's servers, but also satisfies some users' privacy requirements. In yet another embodiment, the user profile may be created and updated on the client **118**, but stored in the user profile server **110**. Such embodiment combines some of the benefits illustrated in the other two embodiments. It is understood by a person of ordinary skill in the art that the user profiles of the present invention can be implemented using client computers, server computers, or both.

[**0049**] **FIG. 3** illustrates an exemplary data structure, a term-based profile table **300**, that may be used for storing term-based profiles for a plurality of users. Table **300** includes a plurality of records **310**, each record corresponding to a user's term-based profile. A term-based profile record **310** includes a plurality of columns including a USER_ID column **320** and multiple columns of (TERM, WEIGHT) pairs **340**. The USER_ID column stores a value that uniquely identifies a user, which may be the USER_ID itself, or a hash thereof. For a given user, there is a set of (TERM, WEIGHT) pairs, where each (TERM, WEIGHT) pair **340** includes a term, typically 1-3 words long, that is usually important to the user, and a weight associated with the term that quantifies the importance of the term. In one embodiment, the term may be represented as one or more n-grams. An n-gram is defined as a sequence of n tokens, where the tokens may be words. For example, the phrase "search engine" is an n-gram of length 2, and the word "search" is an n-gram of length 1. A particular USER_ID may also be used to identify a group of users.

[**0050**] N-grams can be used to represent textual objects as vectors. This makes it possible to apply geometric, statistical and other mathematical techniques, which are well defined for vectors, but not for objects in general. In the present invention, n-grams can be used to define a similarity measure between two terms based on the application of a mathematical function to the vector representations of the terms.